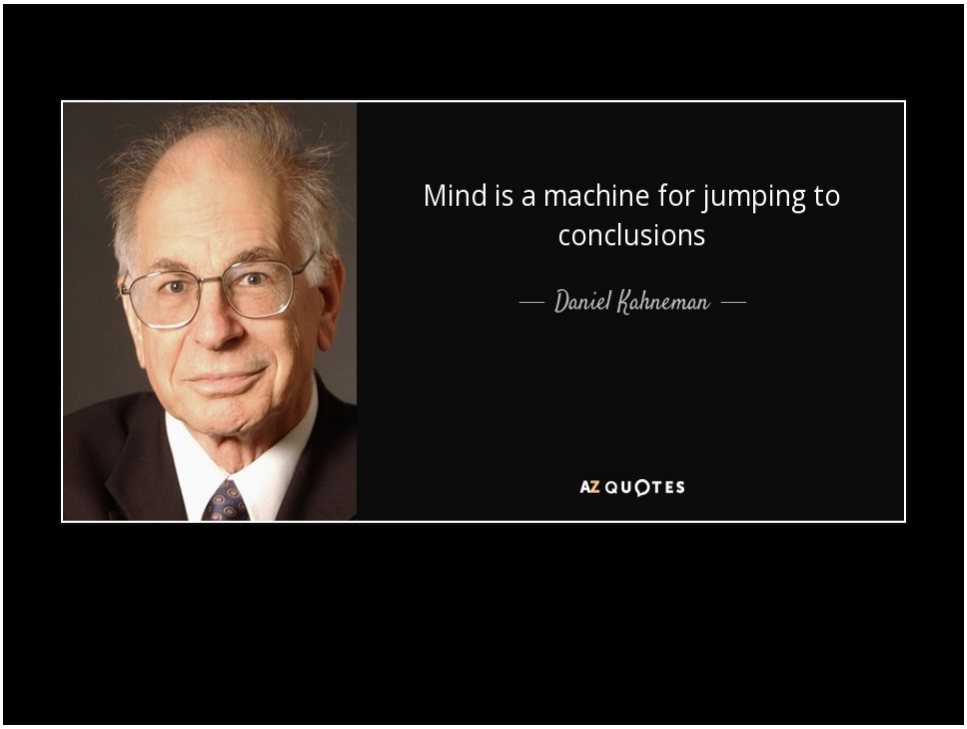
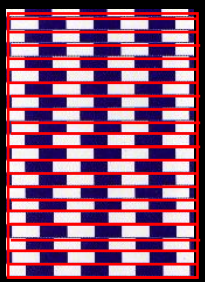

Judgment and decision-making in software engineering
How rational are we?
Magne Jørgensen
Simula Research Laboratory

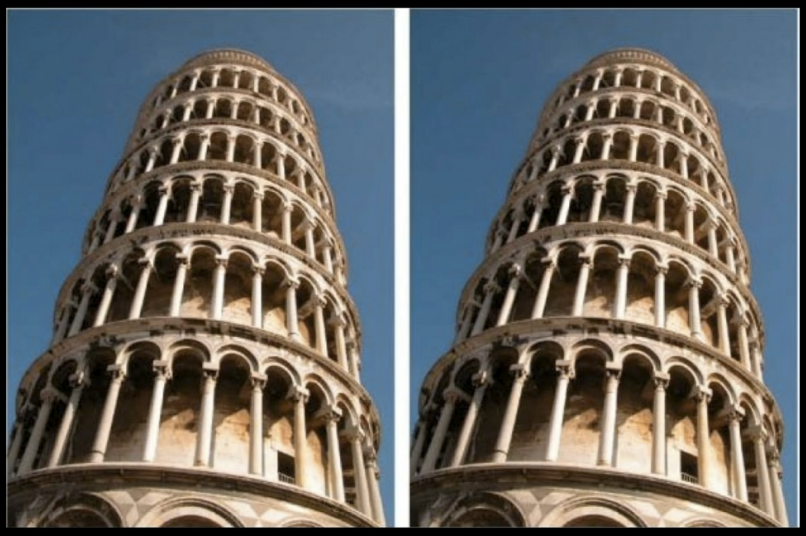

Mind is a machine for jumping to conclusions
— Daniel Kahneman —
AZ QUOTES

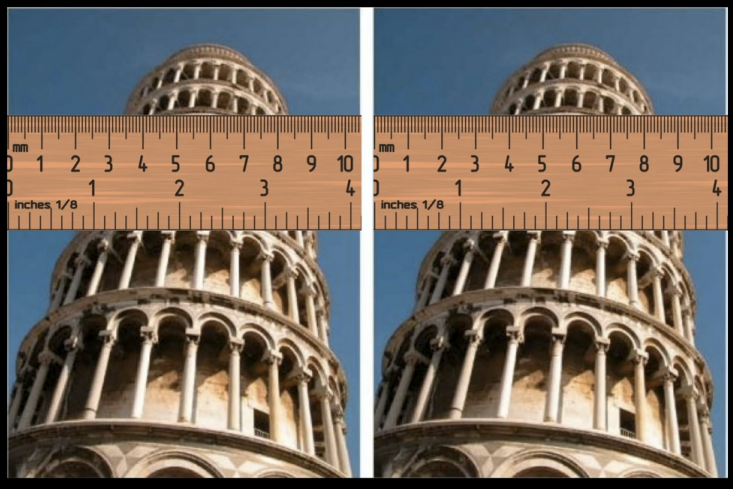
Two systems for thinking (”fast and slow”):
Intuition (expert judgment) and analysis:

Example of the two systems in work
Are these lines parallel?

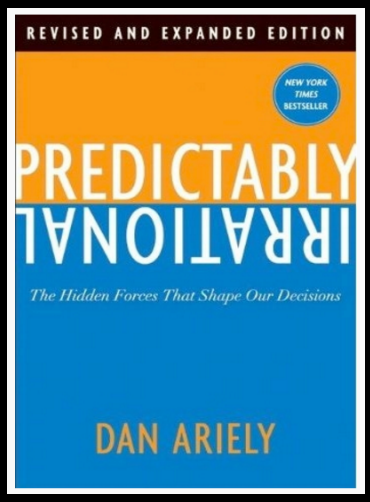


The Leaning Tower of Pisa: Same picture?





OUR JUDGMENTS ARE BIASED - THE BIASES ARE PREDICTABLE



Awarenes of the biases may help

McKinsey-study of 1048 strategic decisions:

- The 25% companies best at avoiding and reducing decision biases had a profit seven times better than the 25% worst.
- Avoiding and reducing decision biases was six times more important for the profit than the amount of or level of detail of analysis preceding the decision.

Dan Lovallo and Olivier Sibony. "The case for behavioral strategy." McKinsey Quarterly 2 (2010): 30-43. www.edpiccolino.com/workspace/articles/mckinsey-the-case-for-behavioral-strategy.pdf)

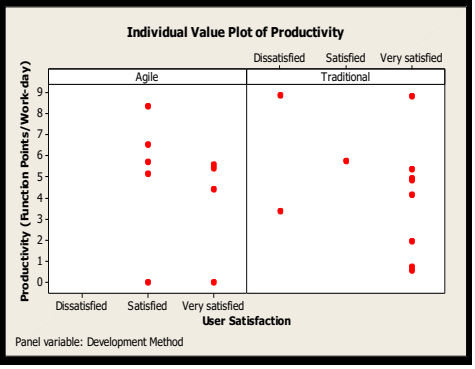
EXAMPLES OF BIASES

Judging is often based on guessing
(seeing what we expect to see)

"I see it when I believe it"

- **Design:**
  - Data sets with randomly set performance data comparing "traditional" and "agile" methods.
  - Survey of each developer's belief in agile methods
- **Question**: How much do you, based on the data set, agree in: *"Use of agile methods has caused a better performance when looking at the combination of productivity and user satisfaction."*
- **Result**: Previous belief in agile determined what they saw in the randomly generated data

Jørgensen, Magne. "Myths and Over-Simplifications in Software Engineering." Lecture Notes on Software Engineering 1.1 (2013): 7.



Focusing on one thing lead us to ignore other things - even gorillas

WHAT YOU SEE IS ALL THERE IS (THE WYSIATI EFFECT)

Focus on identifying risks may lead to more ignorance of what we don't know

Information and Software Technology 52 (2010) 506–516

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Identification of more risks can lead to increased over-optimism of and over-confidence in software development effort estimates

Magne Jørgensen *

Simula Research Laboratory, P.O. Box 134, NO-1325 Lysaker, Norway and University of Oslo, P.O. Box 1072, Blindern, NO-0316 OSLO, Norway

ARTICLE INFO

Article history:
Received 5 July 2009
Received in revised form 7 December 2009
Accepted 8 December 2009
Available online 28 December 2009

Keywords:
Effort estimation
Risk assessment
Human judgment

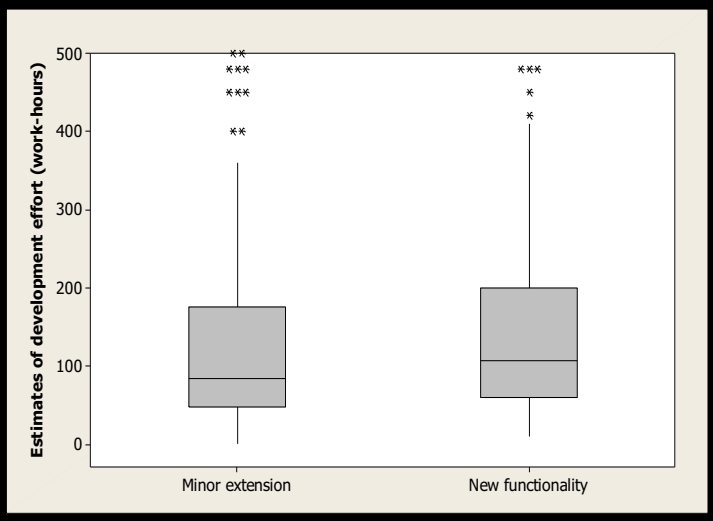
ABSTRACT

Software professionals are, on average, over-optimistic about the required effort usage and over-confident about the accuracy of their effort estimates. A better understanding of the mechanisms leading to the over-optimism and over-confidence may enable better estimation processes and, as a consequence, better managed software development projects. We hypothesize that there are situations where more work on risk identification leads to increased over-optimism and over-confidence in software development effort estimates, instead of the intended improvement of realism. Four experiments with software professionals are conducted to test the hypothesis. All four experiments provide results in support of the hypothesis. Possible explanations of the counter-intuitive finding relate to results from cognitive science on "illusion-of-control", "cognitive accessibility", "the peak-end rule" and "risk as feeling." Thorough work on risk identification is essential for many purposes and our results should not lead to less emphasis on this activity. Our results do, however, suggest that it matters how risk identification and judgment-based effort estimation processes are combined. A simple approach for better combination of risk identification work and effort estimation is suggested.

Framing matters

”20% fat” or ”80% fat free”
Which one would you buy?

Framing in effort estimation



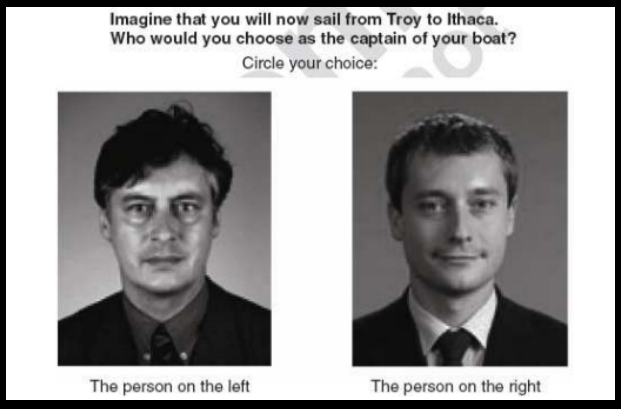
Jorgensen, Magne, and Stein Grimstad. "Avoiding irrelevant and misleading information when estimating development effort." IEEE software 25.3 (2008): 78.

Framing in choice of time unit: Do you think work-hours or workdays gives higher effort estimates?

	Most likely effort (mean)
Estimates in workdays, *n*=29	177 work-hours (std. dev. 254)
Estimates in work-hours, *n*=19	72 work-hours (std. dev. 57)
Difference between Groups	105 work-hours (59% decrease)
T-test of difference (*p*-value)	0.02[1]

Paper accepted for publication in Journal of Systems and Software

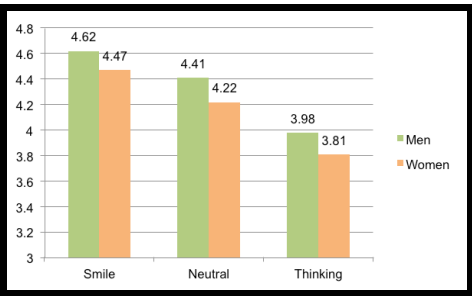
Influence from irrelevant information: Evaluation of people takes less than a second



Election outcome (the winner vs the runner up) in France was predicted by Swiss adults (72% correct, based on rating on perceived competence) and Swiss children aged 5-13 years (71% correct, based on "captain of your boat" selection)!

Antonakis, John, and Olaf Dalgas. "Predicting elections: Child's play!." Science 323.5918 (2009): 1183-1183.

Which pose makes you look most competent?



Petra Filkukova and Magne Jørgensen, The Effect of Facial Expression on Evaluation of Competence: Keep Smiling! Poster at: 17th General Meeting of the European Association of Social Psychology, The Netherlands, Amsterdam., 2014.

Unsupported belief in that that we are good at evaluating peoples competence through interviews

Clouds make nerds look good: field evidence of the impact of incidental factors on decision making

Uri Simonsohn*

Article first published online: 9 OCT 2006

DOI: 10.1002/bdm.545

Copyright © 2006 John Wiley & Sons, Ltd.

Issue

Journal of Behavioral Decision Making

Volume 20, Issue 2, pages 143–152, April 2007

(Incorrect) Trend analyses based on two observations

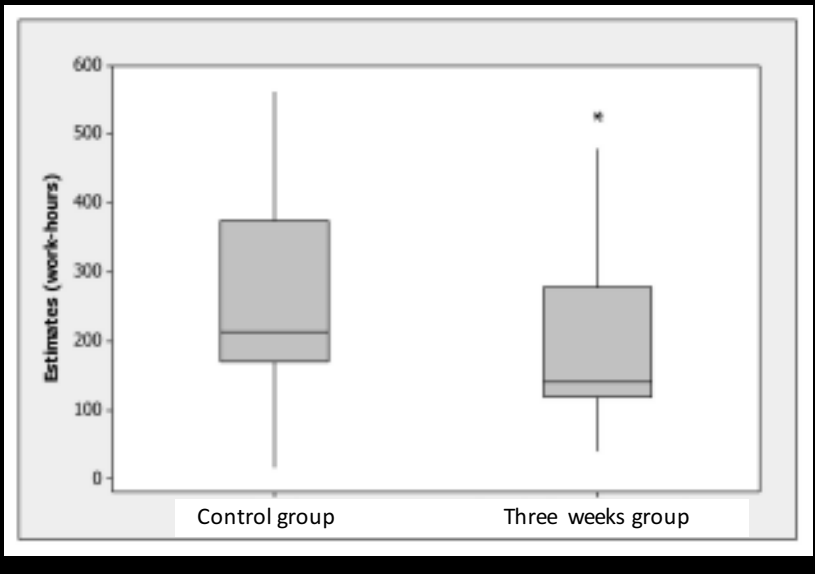
In Study 3, people were told about the landslide risk in three different areas, expressed with colors

Mountain area	Risk 2 weeks ago (T1)		Present risk (T2)		Who will be the most worried?
Lomhesten	Yellow	→	Yellow	→	
Svartoksen	Green	→	Yellow	→	
Buttdalshornet	Red	→	Yellow	→	

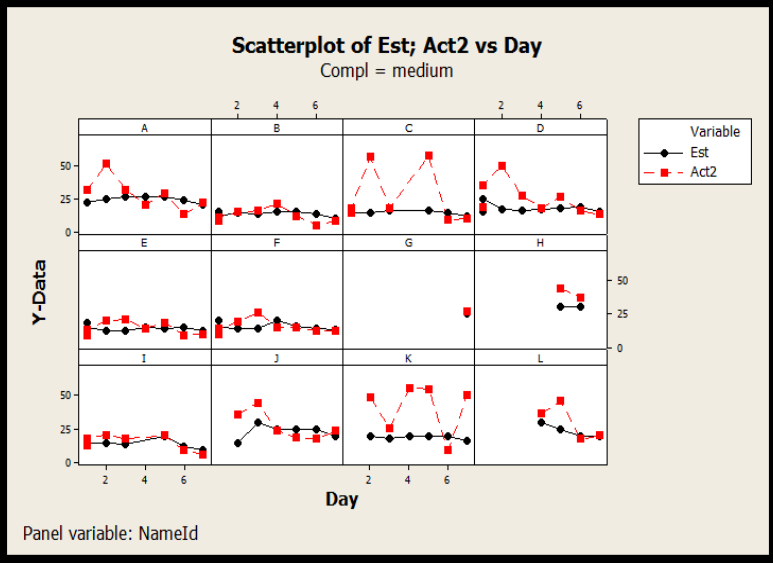
Hohle, Sigrid Møyner, and Karl Halvor Teigen. "Forecasting forecasts: The trend effect." Judgment and Decision Making 10.5 (2015): 416.

Mix of "I believe" and "I wish"

(Companies were asked: "How much effort do you need?" or "It has to be done in three weeks. How much effort do you need?")



"I am my best me" … (and better than average)



We see expertise where there are none (and confuse experience with expertise)

Worse Than Chance? Performance and Confidence Among Professionals and Laypeople in the Stock Market

JOURNAL OF SOFTWARE MAINTENANCE AND EVOLUTION: RESEARCH AND PRACTICE
J. Softw. Maint. Evol.: Res. Pract. 2002; **14**:123–146 (DOI: 10.1002/smr.248)

Research

Impact of experience on maintenance skills

Magne Jørgensen*,† and Dag I. K. Sjøberg

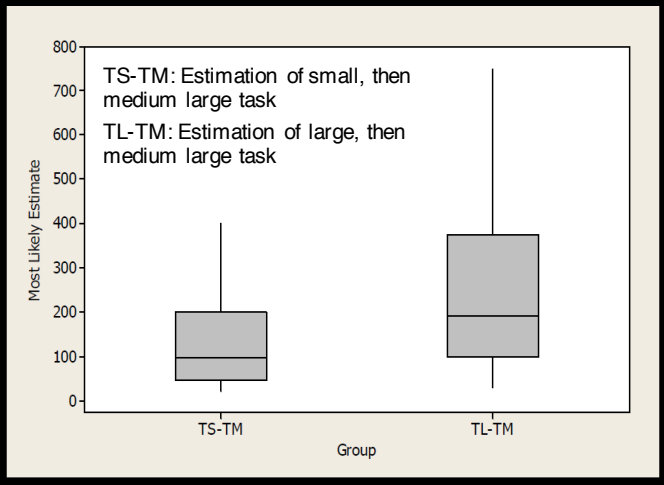
Simula Research Laboratory, Oslo, Norway

We tend to confuse length of experience with expertise/skill!

SUMMARY

This study reports results from an empirical study of 54 software maintainers in the software maintenance department of a Norwegian company. The study addresses the relationship between amount of experience and maintenance skills. The findings were, amongst others, as follows. (1) While there may have been a reduction in the frequency of major unexpected problems from tasks solved by very inexperienced to medium experienced maintainers, additional years of general software maintenance experience did not lead to further reduction. More application specific experience, however, further reduced the frequency of major unexpected problems. (2) The most experienced maintainers did not predict maintenance problems better than maintainers with little or medium experience. (3) A simple one-variable model outperformed the maintainers' predictions of maintenance problems, i.e. the average prediction performance of the maintainers seems poor. An important reason for the weak correlation between length of experience and ability to predict maintenance problems may be the lack of meaningful feedback on the predictions.

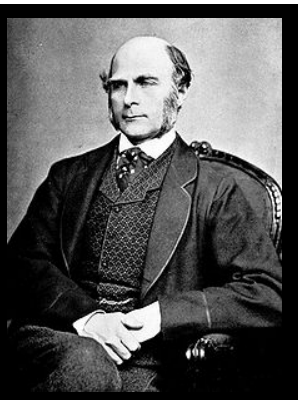
Sequence matters (assimilation effect)



Grimstad, Stein, and Magne Jorgensen. "Preliminary study of sequence effects in judgment-based software development work-effort estimation." Software, IET 3.5 (2009): 435-441.

Example of implication of the assimilation effect (combined with Tversky's feature matching):

The less we know about a new project, the more it looks like the project we chose to compare it with! (e.g., our last project)



We're easily fooled by regression effects (selection bias)
(winner's curse, optimizer's curse, regression towards the mean)

NBA Finals: Spurs hope to break Sports Illustrated cover jinx

W. Scott Bailey
Reporter/Project Coordinator-
San Antonio Business Journal
Email | Twitter | Google+ | Facebook

The national media is showing the San Antonio Spurs some love in advance of the 2013 NBA Finals, which tip off on June 6.

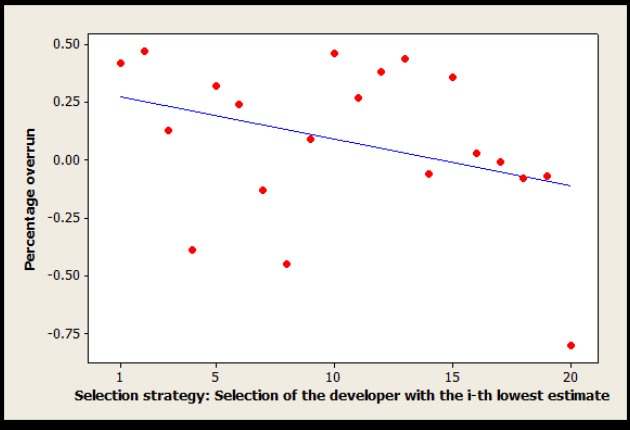
Sports Illustrated has unveiled a cover for its June 10 issue titled: "The Biggest Three."

Sports Illustrated's Chris Ballard's writes in his accompanying story that "it's hard to argue" against proclaiming the Spurs' most talented core — Tim Duncan, Tony Parker and Manu Ginobili — as the most talented trio in NBA history.

Of course, three of the five SI writers who have predicted the outcome of these

Sports Illustrated featured the Spurs' big men, Tim Duncan, Manu Ginobili and Tony Parker on the cover.

We think cost overrun is caused by over-optimism when it's in reality is caused by a selection effect (i.e., by the client's provider selection strategy)



Study:
20 developers estimating and completing the same five tasks

M. Jørgensen. The Influence of Selection Bias on Effort Overruns in Software Development Projects, Information and Software Technology 55(9):1640-1650, 2013.

WHY NOT STUDY BIASES THROUGH INTERVIEWS, SELF-REPORTING, SURVEYS OR THINK-ALOUD PROTOCOL STUDIES

Example: Anchoring effects

Experiment:

- HIGH (LOW) group: *"The customer has indicated that he believes that **1000 (50)** work-hours is a reasonable effort estimate for the specified system. However, the customer knows very little about the implications of his specification on the development effort and you shall not let the customer's expectations impact your estimate. Your task is to provide a realistic effort estimate of a system that meets the requirements specification and has a sufficient quality."*
- Participants: Experienced software developers.
- All (HIGH, LOW, CONTROL group participants) received the same requirement specification.

Jørgensen, Magne, and Dag IK Sjøberg. "The impact of customer expectation on software development effort estimates." International Journal of Project Management 22.4 (2004): 317-325.

Results: Anchoring effects

- Results:
  - HIGH group average: 555 work-hours
  - CONTROL group (no anchor) average: 456 work-hours
  - LOW group average: 99 work-hours!!!
- **None of the developers thought they were much affected by the client expectation.**
- Even when we know that we have been affected by irrelevant or misleading information we don't know how much
- Follow-up experiments give that even when learning about and being warned just before an anchor occur do not remove the anchoring effect

Consequences of the unconscious part of expert judgment for empirical studies in software engineering

- We cannot ask the experts about their mental processes, because they (we) will not know the answers (but will nevertheless answer). Think-aloud protocols does not help much.
- We cannot observe the experts to find out their mental processes. These processes are inside their heads.
- We need to combine theories/results on expert judgment with carefully designed experiments (or natural experiments) in software engineering contexts to find out more about the unconscious (intuition-based) part of expert judgment in software engineering contexts.

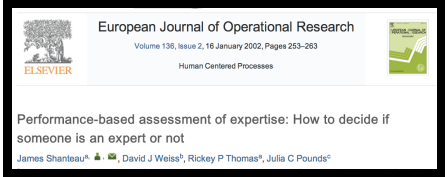
WHAT TO DO TO AVOID JUDGMENT BIASES

KNOW WHEN EXPERT JUDGMENTS ARE LIKELY TO BE GOOD AND WHEN LIKELY TO BE BIASED

Table 2. -- Task characteristics associated with good (left side) and poor (right side) performance in experts.

Characteristics Associated with Good Performance	Characteristics Associated with Poor Performance
Static Stimuli	Dynamic (Changeable) Stimuli
Decisions About Things	Decisions About Behavior
Experts Agree on Stimuli	Experts Disagree on Stimuli
More Predictable Problems	Less Predictable Problems
Some Errors Expected	Few Errors Expected
Repetitive Tasks	Unique Tasks
Feedback Available	Feedback Unavailable
Objective Analysis Available	Subjective Analysis Only
Problem Decomposable	Problem Not Decomposable
Decision Aids Common	Decision Aids Rare

When is an expert a true expert?



INTRODUCE GOOD ANALYTICAL, PROCESSES THAT ARE MORE ROBUST TOWARDS BIASES

Examples of elements of good judgment processes

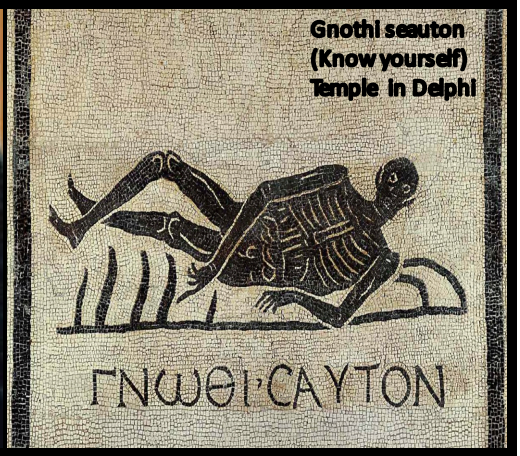
- Get a precise understanding of the problem/decision to make
- Questioning your assumptions (and be open about them)
- Postpone taking the decision before you have collected evidence
- Collect evidence in a balanced and neutral way. Avoid confirmation bias.
- Avoid irrelevant and misleading information
  - Do not think you can get de-biased from it
  - If not avoided, perhaps let someone else make the judgment!
- Combine judgment from independent sources and many perspectives
- Support the judgment with evidence-based methods, especially when in contexts where expert judgment are known to be poor:
  - In project planning, for example, checklists, premortem analysis, review of previous experience may help.
- Reward critique of your judgment and be critical yourself
- Test your intuitions (Dan Arieli)

Examples of judgment process checkpoints

- ”Vested interests”?
- ”In love with the project/idea” biases?
- Groupthink?
- Halo-effect?
- Confirmation bias?
- WYSIATI-bias?
- Anchoring effect
- Sunk cost?
- Over-optimism/not looking back?
- Ignoring worst case?
- Fear of making difficult decisions?

Kahneman, Daniel, Dan Lovallo, and Olivier Sibony. "Before you make that big decision." Harvard business review 89.6 (2011): 50-60.

Summarized:
Think slow and know yourself
(and other people's biases)



”Nudges” affects our behavior (at least for a while)

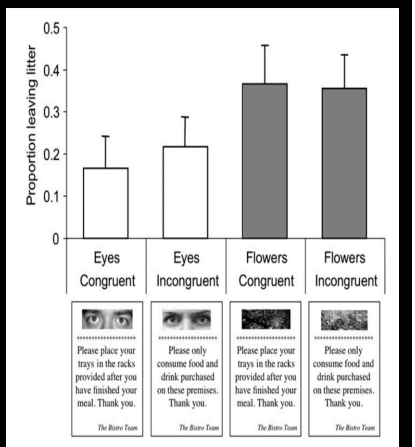
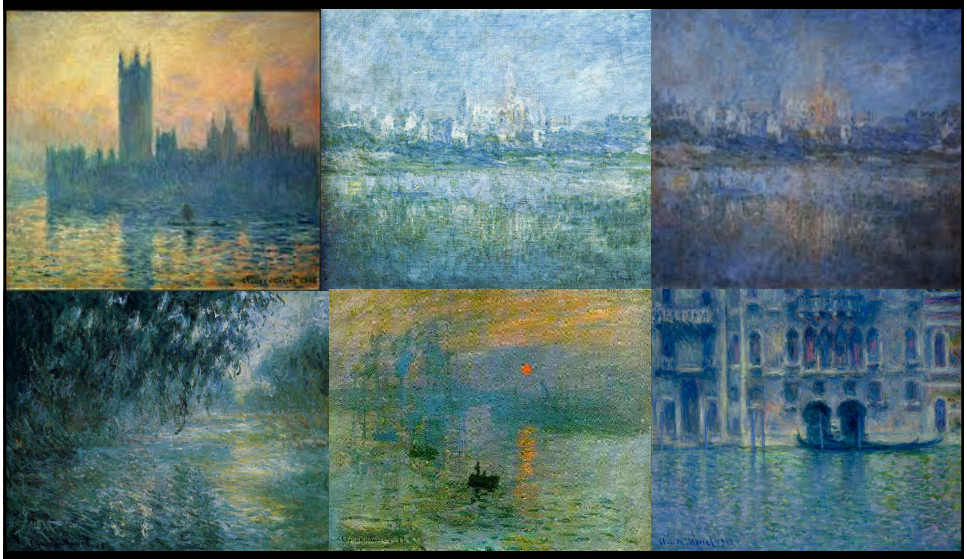


Fig. 1. Proportion of tables leaving litter by condition. An example of one of the eight posters used in each condition is displayed below the graph. Error bars indicate 95% confidence intervals.

When a decision is made, everything looks different (Dan Gilbert)



ted.com/talks/lang/eng/dan_gilbert_asks_why_are_we_happy.html